Machine Learning

Anomaly and Outlier Analysis

S. Nõmm

[1]Department of Software Science, Tallinn University of Technology

13.02.2024

Introduction

1. "An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism."
2. Outliers are also referred to as abnormalities, discordant, deviants, or anomalies in the data mining and statistics literature.
3. Alternatively, one may distinguish the notions of **outlier** and **anomaly** keeping (1) as definition of the outlier and treating anomalies as the sets of outliers which could not be considered as clusters or classes (for example do not satisfy conditions of being a cluster (number of elements, density, etc.))

Applications

- Data cleaning. (NB! Danger to confuse outliers and noise?)
- Model validation and tuning.
- Credit card fraud. Network intrusion detection. and similar Cybersecurity problems.
- Network intrusion detection.

Principles

- Most outlier detection methods create a model of normal patterns.
- Outliers are defined as data points that do not naturally fit within this normal model.
- The "outlierness" of a data point is quantified by a numeric value, known as the outlier score.
  - Real-valued outlier score quantifies the tendency for a data point to be considered an outlier.
  - Binary label is output, indicating whether or not a data point is an outlier.

Models of the normal patterns

- **Extreme values:** A data point is an extreme value, if it lies at one of the two ends of a probability distribution. [Hawkins].
- **Clustering models:** Clustering is considered a complementary problem to outlier analysis.
- **Distance-based models:** In these cases, the k-nearest neighbor distribution of a data point is analyzed to determine whether it is an outlier. Distance-based models can be considered a more fine-grained and instance-centered version of clustering models.
- **Density-based models:** The local density of a data point is used to define its outlier score.
- **Probabilistic models:** The steps are almost analogous to those of clustering algorithms, except that the EM algorithm is used for clustering, and the probabilistic fit values are used to quantify the outlier scores of data points (instead of distance values).
- **Information-theoretic models:** Constrain the maximum deviation allowed from the normal model and then examine the difference in space requirements for constructing a model with or without a specific data point. If the difference is large, then this point is reported as an outlier.

Extreme Value analysis

- Extreme value analysis is a very specific kind of outlier analysis where the data points at the outskirts of the data are reported as outliers. Such outliers correspond to the statistical tails of probability distributions.
- All extreme values are outliers, but the reverse may not be true.
- Consider $\{1, 3, 3, 3, 50, 97, 97, 97, 100\}$.
- 1 and 100 are extreme values and therefore an outliers according to the [Hawkins]. 50 is the mean and is therefore not an extreme value but at the same time it is most isolated point and should be treated as the outlier.

Univariate Extreme Value Analysis

- Univariate extreme value analysis is related to the notion of statistical tail confidence tests. Provides a level of confidence about whether or not a specific data point is an extreme value.
- Most commonly used in the cases of normal distribution. The density function with mean μ and standard deviation σ is

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

- Z - value.

$$z_i = \frac{x_i - \mu}{\sigma}$$

- Large positive values of z_i correspond to the upper tail, whereas large negative values correspond to the lower tail.
- The normal distribution can be directly written in terms of Z - value.

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-z_i^2}{2}}$$

Multivariate Extreme Values

- Defined for unimodal probability distributions with a single peak.
- Let μ be the d-dimensional mean vector of a d-dimensional data set, and Σ be its $d \times d$ covariance matrix.
- The probability density function of d - dimensional data point $\bar{X}$ is

$$f(\bar{X}) = \frac{1}{\sqrt{|\Sigma|(2\pi)^d}} e^{\frac{1}{2}(\bar{X}-\mu)\Sigma^{-1}(\bar{X}-\mu)^T}$$

- Intuitively, this approach models the data distribution along the various uncorrelated directions as statistically independent normal distributions and standardizes them so as to provide each such direction equal importance in the outlier score.

Probabilistic Models

Mixture-based generative model assumes that the data were generated from a mixture of k distributions with the probability distributions $\mathcal{G}_1 \ldots \mathcal{G}_k$ based on the following process:

1. Select a mixture component with prior probability α_i, where $i \in \{1, \ldots, k\}$. Let rth component is selected.
2. Generate data point from $\mathcal{G}_r$.

Denote generative model as $\mathcal{M}$

Probabilistic Models

- The probability density function of the data point $\bar{X}_j$ being generated by the model is :

$$f\left(\bar{X}_j|\mathcal{M}\right) = \sum_{j=1}^{k} \alpha_i f^i\left(\bar{X}_j\right)$$

- For data set $\mathcal{D}$ containing n data points the probability density of the data set being generated by model $\mathcal{M}$ is

$$f\left(\mathcal{D}|\mathcal{M}\right) = \prod_{j=1}^{n} f\left(\bar{X}_i|\mathcal{M}\right)$$

- The log-likelihood fit $\mathcal{L}\left(\mathcal{D}|\mathcal{M}\right)$ of the data set $\mathcal{D}$ with respect to $\mathcal{M}$ is

$$\mathcal{L}\left(\mathcal{D}|\mathcal{M}\right) = \log\Big(\prod_{j=1}^{n} f\left(\bar{X}_j|\mathcal{M}\right)\Big) = \sum_{j=1}^{n} \log\Big(\sum_{i=1}^{k} \alpha_i f^i\left(\bar{X}_j\right)\Big)$$

Clustering for Outlier Detection

- The detection of outliers as a side-product of clustering methods is, however, not an appropriate approach because clustering algorithms are not optimized for outlier detection.
- A simple way of defining the outlier score of a data point is to first cluster the data set and then use the raw distance of the data point to its closest cluster centroid.
- Clustering methods are based on global analysis. Therefore, small, closely related groups of data points will not form their own clusters in most cases.
- The major problem with clustering algorithms is that they are sometimes not able to properly distinguish between a data point that is ambient noise and a data point that is a truly isolated anomaly.

Local Outlier Factor (LOF)

LOF is a normalized density based approach. Denote $V^k(X)$ - distance to its k-nearest neighbour and $L_k(X)$ the set of points within k-nearest neighbor distance of X.

- Reachability distance of X with respect to Y is defined as:

$$R_k(X,Y) = \max\{S(X,Y), V^k(Y)\}$$

- The average reachability distance:

$$AR_k(X) = \frac{\sum_{Y \in L_k(X)} R_k(X,Y)}{|Y \in L_k(X)|}$$

- Local Outlier Factor:

$$LOF_k(X) = \frac{\sum\limits_{Y \in L_k(X)} \frac{AR_k(X)}{AR_k(Y)}}{|Y \in L_k(X)|}$$

Distance-Based Outlier Detection

- The distance-based outlier score of an object O is its distance to its k-th nearest neighbor.
- The distance-based outlier score of an object O the average distance to the k-nearest neighbors.
- Pruning methods are used only for the case where the top-r ranked outliers need to be returned, and the outlier scores of the remaining data points are irrelevant (can be used only for the binary-decision version).
- Local Distance Correction Methods (Local Outlier Factor (LOF) method).
- Histogram- and Grid-Based Techniques

Information-Theoretic Models

- Measure the increase in model size required to describe the data as concisely as possible.
- consider two strings:
  1. ABABABABABABABABABABABABABABABABAB
  2. ABABACABABABABABABABABABABABABABAB

  the first one is the 17 reapeatings of AB but the second one contains one element which makes descriptive model more complicated.
- In general, outliers increase the length of the description in terms of these condensed components to achieve the same level of approximation. For example, a data set with outliers will require a larger number of mixture parameters, clusters, or frequent patterns to achieve the same level of approximation. Therefore, in information-theoretic methods, the components of these summary models are loosely referred to as "code books."